Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Nonparametric Model Calibration Estimation in Survey Sampling

Giorgio E Montanari[a] & M. Giovanna Ranalli[a]

[a] Giorgio E. Montanari is Professor and M. Giovanna Ranalli is Research Assistant , Dipartimento di Scienze Statistiche, Università degli Studi di Perugia, Perugia, Italy.

The authors thank Jean Opsomer, Salvatore Ingrassia, and two anonymous referees for helpful comments and constructive suggestions. This work was supported by a grant from MIUR (COFIN 2002), Italy, and for the application in Section 6.3 by STAR Research Assistance Agreements CR-829095 and CR-829096 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University and Oregon State University. This manuscript has not been formally reviewed by the EPA. The views expressed here are solely those of the authors. The MIUR and EPA do not endorse any products or commercial services mentioned in this report.
Published online: 01 Jan 2012.

Nonparametric Model Calibration Estimation in Survey Sampling

Giorgio E. MONTANARI and M. Giovanna RANALLI

Calibration is commonly used in survey sampling to include auxiliary information at the estimation stage of a population parameter. Calibrating the observation weights on population means (totals) of a set of auxiliary variables implies building weights that when applied to the auxiliaries give exactly their population mean (total). Implicitly, calibration techniques rely on a linear relation between the survey variable and the auxiliary variables. However, when auxiliary information is available for all units in the population, more complex modeling can be handled by means of model calibration; auxiliary variables are used to obtain fitted values of the survey variable for all units in the population, and estimation weights are sought to satisfy calibration constraints on the fitted values population mean, rather than on the auxiliary variables one. In this work we extend model calibration considering more general superpopulation models and use nonparametric methods to obtain the fitted values on which to calibrate. More precisely, we adopt neural network learning and local polynomial smoothing to estimate the functional relationship between the survey variable and the auxiliary variables. Under suitable regularity conditions, the proposed estimators are proven to be design consistent. The moments of the asymptotic distribution are also derived, and a consistent estimator of the variance of each distribution is then proposed. The performance of the proposed estimators for finite-size samples is investigated by means of simulation studies. An application to the assessment of the ecological conditions of streams in the mid-Atlantic highlands in the United States is also carried out.

KEY WORDS: Auxiliary information; Generalized regression estimator; Local polynomials; Model-assisted approach; Neural networks; Nonparametric regression.

1. INTRODUCTION

Availability of auxiliary information to estimate descriptive parameters of a survey variable in a finite population has become fairly common. Census data, administrative registers, previous surveys, and remote sensing provide a wide and growing range of variables that can be used to increase the precision of estimation procedures. A simple way to incorporate known population means (or totals) of auxiliary variables is through ratio and regression estimation. More general situations are handled by generalized regression estimation (Särndal 1980; Särndal, Swensson, and Wretman 1992) and calibration estimation (Deville and Särndal 1992). Those methods have been proposed within a model-assisted approach to inference, where a working model ξ is assumed to describe the relationship between the auxiliary variables and the survey variable. Estimators are sought to have desirable properties, like asymptotic design unbiasedness (i.e., unbiasedness over repeated sampling from the finite population) and design consistency, irrespective of whether the working model is correctly specified or not, and to be particularly efficient if the model holds true.

Nonetheless, all of these techniques refer to essentially a linear regression model for the underlying relationship between the survey and the auxiliary variables. In this framework, concern is mainly with efficient prediction of the values taken by the survey variable in nonsampled units, rather than with interpretation of the relationship between the variable of interest and the auxiliary ones. As a consequence, the introduction of more general models and flexible techniques to obtain predictions seems of great interest. To this purpose, Wu and Sitter (2001) introduced model calibration, where nonlinear parametric regression models and generalized linear regression models are used to obtain model-assisted estimators by generalizing the calibration method of Deville and Särndal (1992).

Further flexibility is also allowed by assuming a nonparametric class of models for ξ. Kernel smoothing was adopted by Kuo (1988) in a model-based approach to inference. Dorfman (1992), Dorfman and Hall (1993), and Chambers, Dorfman, and Wehrly (1993) studied and extended these techniques. Breidt and Opsomer (2000) first considered nonparametric models for ξ within a model-assisted framework and obtained a local polynomial regression estimator as a generalization of the ordinary generalized regression estimator.

Multivariate auxiliary information can be accounted for in the aforementioned proposals. However, the problem of the sparseness of the regressors' values in the design space makes kernel methods and local polynomials inefficient in practice. This problem is known in literature as the *curse of dimensionality*; in high dimensions sample points sparsely populate the space, neighborhoods that contain even a small number of observations have large radii, and most sample points are close to an edge of the space. Local approximators in such a context run into problems (e.g., Friedman 1994). Attempts to handle multivariate auxiliary information make use of recursive covering in a model-based perspective (Di Ciaccio and Montanari 2001) and of generalized additive modeling in a model-assisted framework (Opsomer, Breidt, Moisen, and Kauermann 2003). All of these new methods require knowledge of the value taken by the auxiliary variables for all units in the population. Even though more restrictive, this requirement nowadays can be met whenever information for a population can be combined at the individual level from different sources (e.g., census data, administrative registers, remote sensing).

In this article we combine model calibration estimation with nonparametric methods and propose *nonparametric model calibration estimators* for a finite population mean. Calibration

Giorgio E. Montanari is Professor (E-mail: *giorgio@stat.unipg.it*) and M. Giovanna Ranalli is Research Assistant (E-mail: *giovanna@stat.unipg.it*), Dipartimento di Scienze Statistiche, Università degli Studi di Perugia, Perugia, Italy. The authors thank Jean Opsomer, Salvatore Ingrassia, and two anonymous referees for helpful comments and constructive suggestions. This work was supported by a grant from MIUR (COFIN 2002), Italy, and for the application in Section 6.3 by STAR Research Assistance Agreements CR-829095 and CR-829096 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University and Oregon State University. This manuscript has not been formally reviewed by the EPA. The views expressed here are solely those of the authors. The MIUR and EPA do not endorse any products or commercial services mentioned in this report.

Journal of the American Statistical Association
December 2005, Vol. 100, No. 472, Theory and Methods
DOI 10.1198/016214505000000141

and nonparametric methods have been considered together also by Chambers (1996, 1998) in a model-based context. Here we adopt a model-assisted approach to inference and extend model calibration like that of Wu and Sitter (2001) using nonparametric methods to obtain the fitted values on which to calibrate. More precisely, we consider neural network learning and local polynomial smoothing to estimate the functional relationship between the survey variable and the auxiliary variables. Opsomer, Moisen, and Kim (2001) sketched this idea as a way to produce survey weights from a nonlinear generalized additive model fit. Although Nordbotten (1996) used neural networks for imputation with auxiliary information coming from administrative registers, the use of neural networks for model calibration is new and allows for more flexible prediction and straightforward insertion of multivariate auxiliary information.

In principle, any nonparametric method existing in the literature can be used to recover fitted values for the survey variable on nonsampled units. However, the treatment here is limited to neural networks and local polynomials as methods for which theoretical properties for the resulting estimators can be established. Moreover, for local polynomials, an existing methodology in the same framework is available, whereas neural networks are widely used in practice, and software is commonly available that can easily handle multivariate data.

The treatment proceeds as follows. In Section 2 we briefly review calibration and the generalized regression estimation technique. Then we introduce the neural network model calibration estimator in Section 3. We state the design theoretical properties of this estimator in Section 4. In Section 5 we introduce the local polynomial model calibration estimator and establish its theoretical properties. In Sections 6.1 and 6.2 we report the results of simulation experiments carried on to study the finite-sample performance of the proposed estimators, and in Section 6.3 we consider nonparametric model calibration for assessment of the ecological condition of streams in the mid-Atlantic highlands. We give some concluding remarks in Section 7.

2. CALIBRATION TECHNIQUES AND REGRESSION

Consider a finite population $\mathcal{U} = \{1, \ldots, N\}$. For each unit in the population, we assume that the value of a vector $\mathbf{x}$ of Q auxiliary variables is available (e.g., from census data, administrative registers, previous surveys, or remote sensing); hence the vector $\mathbf{x}_i = (x_{1i}, \ldots, x_{qi}, \ldots, x_{Qi})$, is known $\forall\, i \in \mathcal{U}$. A sample s of size n is drawn without replacement from $\mathcal{U}$ according to a probabilistic sampling plan with inclusion probabilities π_i and π_{ij}, for all $i, j \in \mathcal{U}$. Let $\delta_i = 1$ when $i \in s$ and $\delta_i = 0$ otherwise; then we have that $E(\delta_i) = \pi_i$, where expectation is taken with respect to the sampling design. The survey variable y is observed for each unit in the sample, and hence y_i is known for all $i \in s$. The goal is to estimate the population mean of the survey variable, that is, $\bar{Y} = N^{-1} \sum_{i=1}^{N} y_i$. For ease of notation in what follows we assume that the design is such that $\sum_{i=1}^{n} \pi_i^{-1} = N$; generalization to cases for which the latter does not hold is straightforward and available from the authors.

Deville and Särndal (1992) first introduced the notion of a calibration estimator. This is defined to be a linear combination of observations $\hat{\bar{Y}}_c = \sum_{i=1}^{n} w_i y_i$ with weights chosen to minimize an average distance from the basic design weights $d_i = 1/\pi_i$. Minimization is constrained to satisfy a set of calibration equations, $N^{-1} \sum_{i=1}^{n} w_i \mathbf{x}_i = \bar{\mathbf{x}}$, where $\bar{\mathbf{x}}$ is the known vector of population means for the auxiliary variables. Although alternative distance measures are available from Deville and Särndal (1992), all resulting estimators are asymptotically equivalent to the one obtained from minimizing the chi-squared distance function

$$\Phi_s = \sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i q_i}, \tag{1}$$

where q_i's are known positive weights unrelated to d_i. This choice provides the following calibration estimator as the solution to the minimization problem:

$$\hat{\bar{Y}}_c = \hat{\bar{Y}} + (\bar{\mathbf{x}} - \hat{\bar{\mathbf{x}}})' \hat{\boldsymbol{\beta}}, \tag{2}$$

where $\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{n} d_i q_i \mathbf{x}_i \mathbf{x}_i')^{-1} \sum_{i=1}^{n} d_i q_i \mathbf{x}_i y_i$, and $\hat{\bar{Y}} = N^{-1} \times \sum_{i=1}^{n} d_i y_i$ and $\hat{\bar{\mathbf{x}}} = N^{-1} \sum_{i=1}^{n} d_i \mathbf{x}_i$ are the Horvitz–Thompson estimators of $\bar{Y}$ and $\bar{\mathbf{x}}$. This definition of $\hat{\bar{Y}}_c$ is equivalent to a generalized regression estimator, which is derived as a model-assisted estimator assuming a linear regression model, with variance structure given by a diagonal matrix with elements $1/q_i$ (Deville and Särndal 1992, sec. 1). Examples of the role of the constants q_i have also been given by Deville and Särndal (1992) and Särndal (1996). Hence $\hat{\bar{Y}}_c$ implicitly relies on a linear relationship between the auxiliary variables and the survey variable. By noting that "it is the relationship between y and $\mathbf{x}$, hopefully captured by the working model, that determines how the auxiliary information should best be used," Wu and Sitter (2001) proposed to generalize the calibration procedure by means of *model calibration*. In particular, they considered generalized linear models and nonlinear regression models for ξ such that $E_\xi(y_i) = \mu(\mathbf{x}_i, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is an unknown superpopulation parameter vector, $\mu(\cdot)$ is a known function of $\mathbf{x}_i$ and $\boldsymbol{\theta}$, and E_ξ denotes expectation with respect to ξ. The proposed model calibration estimator for $\bar{Y}$ is $\hat{\bar{Y}}_{\mathrm{mc}} = N^{-1} \sum_{i=1}^{n} w_i y_i$, with weights again sought to minimize the distance measure Φ_s in (1) under the new constraints $\sum_{i=1}^{n} w_i = N$ and $N^{-1} \sum_{i=1}^{n} w_i \hat{\mu}_i = N^{-1} \sum_{i=1}^{N} \hat{\mu}_i$, where $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\theta}}$ is a design-consistent estimator for $\boldsymbol{\theta}$. In this context, calibration is performed with respect to the population mean of the fitted values $\hat{\mu}_i$, instead of the population mean of the auxiliary variables as for $\hat{\bar{Y}}_c$. The resulting estimator can be written as

$$\hat{\bar{Y}}_{\mathrm{mc}} = \hat{\bar{Y}} + \frac{1}{N} \left\{ \sum_{i=1}^{N} \hat{\mu}_i - \sum_{i=1}^{n} d_i \hat{\mu}_i \right\} \hat{\beta}_{\mathrm{mc}}, \tag{3}$$

where $\hat{\beta}_{\mathrm{mc}} = \sum_{i \in s} d_i q_i (\hat{\mu}_i - \breve{\mu})(y_i - \breve{y}) / \sum_{i \in s} d_i q_i (\hat{\mu}_i - \breve{\mu})^2$, $\breve{y} = \sum_{i \in s} d_i q_i y_i / \sum_{i \in s} d_i q_i$, and $\breve{\mu} = \sum_{i \in s} d_i q_i \hat{\mu}_i / \sum_{i \in s} d_i q_i$. Wu (2003) showed that the resulting estimator is optimal among the class of calibration estimators, in that the expected value of the asymptotic design variance under the model and any regular sampling design with fixed sample size reaches its minimum.

Following another direction to allow for more complex modeling than linear models, Breidt and Opsomer (2000) proposed a model-assisted nonparametric regression estimator based on local polynomial smoothing. It is based on a nonparametric superpopulation model ξ for which

$$y_i = m(x_i) + \varepsilon_i \quad \text{for } i = 1, 2, \ldots, N, \tag{4}$$

where $m(\cdot)$ is a smooth function of a single auxiliary variable x, the ε_i's are independent random variables with mean 0 and variance $v(x_i)$, and $v(\cdot)$ is smooth and strictly positive. A local polynomial kernel estimator of degree p is used to obtain fitted values. Let $K_h(u) = h^{-1}K(u/h)$, where K denotes a continuous kernel function and h is the bandwidth. The role of h on the efficiency of the final estimator is discussed in Section 6.4. Then a sample-based consistent estimator of the local polynomial estimator for the unknown $m(x_i)$ is given by

$$\hat{m}_i = \mathbf{e}_1'(\mathbf{X}_{si}'\mathbf{W}_{si}\mathbf{X}_{si})^{-1}\mathbf{X}_{si}'\mathbf{W}_{si}\mathbf{y}_s, \tag{5}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)'$ is a column vector of length $p+1$, $\mathbf{y}_s = (y_1, \ldots, y_n)'$, $\mathbf{W}_{si} = \operatorname{diag}\{d_j K_h(x_j - x_i)\}_{j\in s}$, and $\mathbf{X}_{si} = [\,1 \quad x_j - x_i \quad \cdots \quad (x_j - x_i)^p\,]_{j\in s}$. The local polynomial regression estimator for $\bar{Y}$ is given by

$$\hat{\bar{Y}}_{lp} = \hat{\bar{Y}} + \frac{1}{N}\left\{\sum_{i=1}^{N}\hat{m}_i - \sum_{i=1}^{n} d_i\hat{m}_i\right\}. \tag{6}$$

Among other desirable properties, the estimator (6) has been proven to be calibrated with respect to the auxiliary variables (Breidt and Opsomer 2000, sec. 2), whereas it is not calibrated with respect to the fitted values $\hat{m}_i$. Moreover, as noted in Section 1, accounting for more than one auxiliary variable could represent a problem in practice. In contrast, model calibration estimators proposed by Wu and Sitter (2001) rely on classes of superpopulation models that could be usefully enlarged to account for more complex model structures. In the following sections we introduce two nonparametric model calibration estimators of the population mean.

3. A NEURAL NETWORK MODEL CALIBRATION ESTIMATOR

Neural networks are very popular learning methods. Among others, Ripley (1996), Hastie, Tibshirani, and Friedman (2001), and Ingrassia and Davino (2002) have shown that this technique is suitable to a wide range of problems. Theoretical work by Cybenko (1989), Funahashi (1989), and Barron (1993) has provided evidence of their universal approximation property of continuous functions. We use a feedforward neural network with skip-layer connections, whose typical structure is represented in Figure 1. Three components are present in such a model: inputs, output, and an intermediate set of hidden variables—*neurons*—that transform in a nonlinear fashion the information coming from the inputs to the output. The three sets of variables are linked only by one-way connections, whose direction is indicated by the arrows. No feedback is allowed, the three layers are totally connected, and there is no link between units belonging to the same layer. Skip-layer connections link straightforwardly the input variables to the output. Each connection is weighted. A linear combination of the inputs is the input to each hidden unit; at this level a constant is added, and an activation function $\phi(\cdot)$ is applied to get outgoing signals to the output. To a linear combination of these signals, another constant is added to provide the final output. Denoting with $f(\mathbf{x}_i)$ the output, the foregoing structure can be formalized as

$$f(\mathbf{x}_i) = \sum_{q=1}^{Q}\beta_q x_{qi} + \sum_{m=1}^{M} a_m \phi\left(\sum_{q=1}^{Q}\gamma_{qm}x_{qi} + \gamma_{0m}\right) + a_0, \tag{7}$$

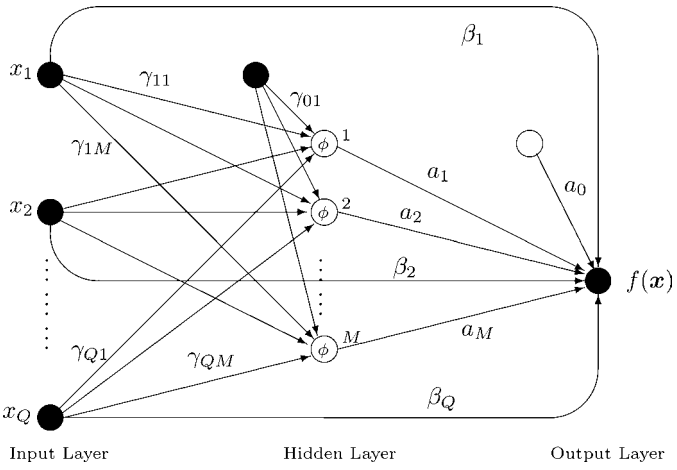

Figure 1. Schematic of a Single Hidden Layer Feedforward Neural Network With Skip Layer Connections.

where M is the number of neurons at the hidden layer; $a_m \in \mathcal{R}$ for $m = 1, \ldots, M$ is the weight of the connection of the mth hidden node with the output; and $\gamma_{qm} \in \mathcal{R}$ for $m = 1, \ldots, M$ and $q = 1, \ldots, Q$ is the weight attached to the connection between the qth input and the mth hidden node. Scalars a_0 and γ_{0m} for $m = 1, \ldots, M$ represent the *activation levels* of the output and the M neurons. The activation function $\phi(\cdot)$ is usually a sigmoidal function, an S-shaped function that assumes monotonically increasing values between 0 and 1 as the value of its argument goes from $-\infty$ to $+\infty$. Finally, by allowing skip-layer connections from the inputs to the output, β_q for $q = 1, \ldots, Q$ denotes the weight attached to each direct connection. Overall, to a basic linear structure provided by the skip-layer connections, nonlinear components are added to enable fitting of more complex regression functions (see, e.g., Ripley 1996, chap. 5). Although feedforward networks with more than one layer of hidden units and more complicated networks that allow feedback of information can be specified, for the sake of simplicity we deal only with the presented structure, which is commonly used for a wide variety of applications and has the appealing feature of being easily implemented using the nnet( ) function in R and S–PLUS.

Now, going back to the issue of estimating $\bar{Y}$, let us assume that the relationship between the survey variable and the auxiliary variables can be described by the following superpopulation model:

$$\begin{cases} E_\xi(y_i) = f(\mathbf{x}_i) & \text{for } i = 1, \ldots, N \\ V_\xi(y_i) = v(\mathbf{x}_i) & \text{for } i = 1, \ldots, N \\ C_\xi(y_i, y_j) = 0 & \text{for } i \neq j, \end{cases} \tag{8}$$

where V_ξ and C_ξ denote variance and covariance, with respect to ξ, $f(\mathbf{x}_i)$ is function (7), and $v(\cdot)$ is smooth and strictly positive. Assuming M to be fixed, we denote by $\boldsymbol{\theta}$ the set of all parameters of the network and write

$$\boldsymbol{\theta} = (\beta_1, \ldots, \beta_Q, a_0, a_1, \ldots, a_M, \gamma_{01}, \ldots, \gamma_{0M}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_M)', \tag{9}$$

where $\boldsymbol{\gamma}_m = (\gamma_{1m}, \ldots, \gamma_{Qm})$ for $m = 1, \ldots, M$. Then $f(\mathbf{x}_i)$ in (8) becomes $f(\mathbf{x}_i; \boldsymbol{\theta})$, and $\boldsymbol{\theta}$ is a vector of unknown superpopulation parameters. Let $\boldsymbol{\theta}^*$ denote the unknown true value of $\boldsymbol{\theta}$.

Remark 1. Model assumptions in (8) formally restrict the regression function to belong to a specified class of nonlinear parametric functions. Nevertheless, the universal approximation property of neural networks proved by Cybenko (1989) and Funahashi (1989) shows that any continuous function can be uniformly approximated on compact sets (i.e., closed and bounded subsets of $\mathcal{R}^Q$) by increasing the size of the hidden layer M in (7). Barron (1993) proved that the approximation error for a fixed M is bounded by a term of order $O(1/M)$ for all functions having a Fourier representation. A detailed review of the theoretical properties of feedforward neural networks was given by Ripley (1996).

To estimate the regression function (7), we follow the approach of Wu and Sitter (2001); that is, we need to define a design-consistent estimate of $\boldsymbol{\theta}^*$ and thus of the regression function at $\mathbf{x}_i$, for $i = 1, \ldots, N$. To that purpose, we first seek an estimate $\tilde{\boldsymbol{\theta}}$ of the model parameter $\boldsymbol{\theta}^*$ based on the entire finite population. Let us define the population parameter $\tilde{\boldsymbol{\theta}}$ as the minimizer in the parameter space Θ of the weighted sum of squared residuals with a weight decay penalty term, that is,

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \left\{ \sum_{i=1}^{N} \frac{1}{v_i}(y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2 + \lambda \sum_{l=1}^{r} \theta_l^2 \right\}, \tag{10}$$

where v_i for $i = 1, \ldots, N$ are known positive weights assumed to be proportional to the variance function $v(\mathbf{x}_i)$, r is the dimension of the vector $\boldsymbol{\theta}$, and λ is a tuning parameter. The weight decay penalty is analogous to ridge regression introduced for linear models as a solution to collinearity. Larger values of λ tend to favor approximations corresponding to small values of the parameters and thus shrink the weights toward 0 to avoid overfitting. Here we assume that M and λ have been fixed in advance to hopefully provide a good fit at the population level. The issue of selecting values for λ and M is discussed in Section 6.4. Then $\tilde{\boldsymbol{\theta}}$ is obtained as the solution of the following equations:

$$\sum_{i=1}^{N} \left\{ (y_i - f(\mathbf{x}_i, \boldsymbol{\theta})) \frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{1}{v_i} - \frac{\lambda}{N}\boldsymbol{\theta} \right\} = \mathbf{0}. \tag{11}$$

The sum on the left side of (11) is a population total; then a design consistent estimate $\hat{\boldsymbol{\theta}}$ of $\tilde{\boldsymbol{\theta}}$ is defined as the solution of the design-based sample version of (11), that is, $\hat{\boldsymbol{\theta}}$ is the solution to

$$\sum_{i=1}^{n} d_i \left\{ (y_i - f(\mathbf{x}_i, \boldsymbol{\theta})) \frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{1}{v_i} - \frac{\lambda}{N}\boldsymbol{\theta} \right\} = \mathbf{0}. \tag{12}$$

Lemma A.1 in the Appendix shows that $\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}} + O_p(n^{-1/2})$, where the order statement is considered with respect to the design. Once the estimates $\hat{\boldsymbol{\theta}}$ are obtained, the available auxiliary information is included in the estimator through the fitted values $\hat{f}_i = f(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$, for $i = 1, \ldots, N$. Toward this end, we define the *neural network model calibration estimator* of $\bar{Y}$ as $\hat{\bar{Y}}_{nn}^{\mathrm{mc}} = N^{-1}\sum_{i=1}^{n} w_i y_i$, where the calibrated weights w_i are sought to minimize the distance function Φ_s in (1) under the constraints $N^{-1}\sum_{i=1}^{n} w_i = 1$ and $N^{-1}\sum_{i=1}^{n} w_i \hat{f}_i = N^{-1}\sum_{i=1}^{N} \hat{f}_i$. The proposed estimator follows to be

$$\hat{\bar{Y}}_{nn}^{\mathrm{mc}} = \hat{\bar{Y}} + \frac{1}{N}\left\{ \sum_{i=1}^{N} \hat{f}_i - \sum_{i=1}^{n} d_i \hat{f}_i \right\} \hat{\beta}_{nn}, \tag{13}$$

where

$$\hat{\beta}_{nn} = \frac{\sum_{i=1}^{n} d_i q_i (\hat{f}_i - \breve{f})(y_i - \breve{y})}{\sum_{i=1}^{n} d_i q_i (\hat{f}_i - \breve{f})^2} \tag{14}$$

and $\breve{f} = \sum_{i=1}^{n} d_i q_i \hat{f}_i / \sum_{i=1}^{n} d_i q_i$.

Estimator (13) mimics estimator $\hat{\bar{Y}}_{\mathrm{mc}}$ in (3) and includes a straightforward extension to neural networks of estimator (6) proposed by Breidt and Opsomer (2000) by setting $\hat{\beta}_{nn} = 1$. Here we add the supplementary regression step performed with $\hat{\beta}_{nn}$. In fact, $\hat{\bar{Y}}_{nn}^{\mathrm{mc}}$ can be seen as a generalized regression estimator based on the working model $E_\xi(y_i) = \alpha + \beta f(\mathbf{x}_i)$. Hence $\hat{\bar{Y}}_{nn}^{\mathrm{mc}}$ uses estimates of $f(\mathbf{x}_i)$ as the auxiliary variable in a generalized regression procedure. Sample-based fits $\hat{f}_i$ are design-consistent estimates of population fits $\tilde{f}_i = f(\mathbf{x}_i, \tilde{\boldsymbol{\theta}})$ (Lemma A.2 in the App.). If the nonparametric technique provides model-unbiased estimates of the mean function $f(\mathbf{x})$ at the population level, then this supplementary calibration step would not provide gains in efficiency with respect to setting the value of $\hat{\beta}_{nn}$ equal to 1. However, in cases in which the nonparametric technique provides biased estimates of the mean function or the working model is not valid, then this step makes sense in a model-assisted approach and will asymptotically lead to more efficient estimates for the population mean of y.

We assess properties of estimator in (13) in the next section. Here we note that the presence of weights d_i in the least squares procedure in (12) makes $\hat{f}_i$ a design-consistent estimator of the population fit $\tilde{f}_i$. The latter is a finite population parameter in as far as the number of hidden units M and weight decay parameter λ are given and fixed. This procedure of deriving fitted values mirrors that used for the development of the generalized regression estimator, to which the proposed estimator reverts as both the number of units in the hidden layer M and the value of λ go to 0.

4. ASSUMPTIONS AND PROPERTIES OF $\hat{\bar{Y}}_{nn}^{\mathrm{mc}}$

To study the design properties of $\hat{\bar{Y}}_{nn}^{\mathrm{mc}}$, we use Taylor series approximations of the fitted values $\hat{f}_i$. Toward this end, we need a set of regularity conditions on the behavior of $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ and of the function $f(\cdot)$ in the asymptotic framework. We assume that there is a sequence of finite populations indexed by ν and a corresponding sequence of sampling designs. Both the population size N_ν and the sample size n_ν approach infinity as $\nu \to \infty$. More details for the asymptotic framework have been given by Isaki and Fuller (1982). We drop the subscript "ν" for ease of notation. To prove our theoretical results, we make the following assumptions:

(a) For each ν, the $\mathbf{x}_i$'s are iid from an unknown and fixed distribution $F(\mathbf{x}) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_Q} g(t_1, t_2, \ldots, t_Q)\, d\mathbf{t}$, where $g(\cdot)$ is a strictly positive density whose support is a compact subset of $\mathcal{R}^Q$.

(b) For each ν, conditioning on the values $\mathbf{x}_i$, the superpopulation model is as in (8). Hence the $\mathbf{x}_i$'s are considered fixed with respect to the superpopulation model ξ.

(c) The survey variable has bounded fourth moment with ξ-probability 1.

(d) The sampling rate is bounded, that is,

$$\limsup_{\nu\to\infty} nN^{-1} = \pi,$$

where $\pi \in (0, 1)$.

(e) For any study variable z with bounded fourth moment, the sampling design $p(s)$ is such that the Horvitz–Thompson estimator of the population mean $\bar{Z}$ is asymptotically normally distributed and is design-consistent with variance $O(n^{-1})$; the latter can be consistently estimated by the Horvitz–Thompson variance estimator.

(f) The parameter space Θ is a compact set, and $\boldsymbol{\theta}^*$ is an interior point of Θ and is irreducible; that is, for $m, m' \neq 0$, none of the following three cases holds (Hwang and Ding 1997): $a_m = 0$ for some $m = 1, \ldots, M$; $\boldsymbol{\gamma}_m = \mathbf{0}$ for some $m = 1, \ldots, M$; and $(\boldsymbol{\gamma}_m, \gamma_{0m}) = \pm(\boldsymbol{\gamma}_{m'}, \gamma_{0m'})$ for some $m \neq m'$.

(g) The activation function ϕ in (7) is a symmetric sigmoidal function differentiable to any order; moreover, the class of functions $\{\phi(bt + b_0), b > 0\} \cup \{\phi \equiv 1\}$ is assumed to be linearly independent. The logistic activation function $\phi(t) = [1 + \exp(-t)]^{-1}$ fulfills these requirements; other examples of sigmoidal functions satisfying these conditions have been given by Hwang and Ding (1997).

Remark 2. Sufficient conditions for the existence of a sampling design as in assumption (e) have been given by, for example, Fuller (1975), Fuller and Isaki (1981), Kott (1990), and Thompson (1997, chap. 3).

Remark 3. Assumptions (f) and (g) concern the neural network structure and allow identifiability of the network parameters to some extent. In fact, every neural network is unidentifiable, in the sense that there are transformations on the parameter vector $\boldsymbol{\theta}$ that leave $f(\mathbf{x}; \boldsymbol{\theta})$ invariant. Nonetheless, if we rewrite (9) as $\boldsymbol{\theta} = \{\beta_1, \ldots, \beta_Q, a_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_M\}$, where $\boldsymbol{\alpha}_m = (a_m, \gamma_{0m}, \boldsymbol{\gamma}_m)$ for $m = 1, \ldots, M$, then, under assumptions (f) and (g) there are only two kinds of transformations that leave $f(\mathbf{x}; \boldsymbol{\theta})$ invariant (Hwang and Ding 1997, thm. 2.3):

- *Permutation*: The function is unchanged if we permute $\boldsymbol{\alpha}_m$'s
- *Sign flips*: Because the activation function is odd, $a_m \times \phi(\boldsymbol{\gamma}_m \mathbf{x} + \gamma_{0m}) = a_m - a_m \phi(-\boldsymbol{\gamma}_m \mathbf{x} - \gamma_{0m})$, and hence the pair of parameters $(a_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m, \ldots, \boldsymbol{\alpha}_M)$ and $(a_0 + a_m, \boldsymbol{\alpha}_1, \ldots, -\boldsymbol{\alpha}_m, \ldots, \boldsymbol{\alpha}_M)$ give exactly the same value of $f(\mathbf{x}; \boldsymbol{\theta})$.

These two transformations generate a family of $2^M M!$ elements. For all transformations τ in this family, it is $f(\mathbf{x}; \boldsymbol{\theta}) = f(\mathbf{x}; \tau(\boldsymbol{\theta}))$. Each transformation can be characterized as being a composite function of $\{\tau_1, \ldots, \tau_M\}$, where

$$\begin{aligned}
&\tau_1(a_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_M) \\
&\quad = (a_0 + a_1, -\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_M), \\
&\tau_m(a_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_M) \\
&\quad = (a_0, \boldsymbol{\alpha}_m, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{m-1}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_{m+1}, \ldots, \boldsymbol{\alpha}_M) \\
&\qquad \text{for } m = 2, \ldots, M.
\end{aligned} \tag{15}$$

Thus assumptions (f) and (g) allow $\boldsymbol{\theta}$ to be identifiable up to the family of transformations generated by (15). That is, if there exists another $\breve{\boldsymbol{\theta}}$ such that $f(\mathbf{x}; \breve{\boldsymbol{\theta}}) = f(\mathbf{x}; \boldsymbol{\theta})$, then there exists a transformation generated by (15) that transforms $\breve{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}$. This allows for overcoming identifiability problems by constructing parameter subspaces within which $\boldsymbol{\theta}$ is identifiable. Hwang and Ding (1997) proposed such a construction when the parameters are estimated without weight decay. Extension of their method to situations in which $\lambda \neq 0$ is straightforward, because for sufficiently large N, all of the minimizers of (10) tend to be the same as those of $\sum_{i=1}^N (y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2 / v(\mathbf{x}_i)$.

Let T_i for $i = 1, \ldots, k$, where $k = 2^M M!$ be the transformations generated by (15). Then let $\boldsymbol{\theta}_i^* = T_i(\boldsymbol{\theta}^*)$, for $i = 1, \ldots, k$, be all of the transformations of the true parameter $\boldsymbol{\theta}^*$. Because $\boldsymbol{\theta}^*$ is irreducible, they are all distinct. Therefore, balls $\mathcal{B}(\boldsymbol{\theta}_i^*, r_i)$ centered at $\boldsymbol{\theta}_i^*$ with radius $r_i > 0$ may be chosen to be disjoint. For sufficiently large N, all of the least squares estimates $\{\tilde{\boldsymbol{\theta}}_i\}$ will be in $\mathcal{B} = \bigcup_{i=1}^k \mathcal{B}(\boldsymbol{\theta}_i^*; r_i)$, with ξ-probability 1 (Hwang and Ding 1997). Therefore, $\mathcal{B}$ can be assumed to be the parameter space without loss of generality, and each ball $\mathcal{B}(\boldsymbol{\theta}_i^*, r_i)$ will be a subset Θ_i of the parameter space. If we restrict to $\Theta_1 = \mathcal{B}(\boldsymbol{\theta}_1^*; r_1)$, then $\tilde{\boldsymbol{\theta}}_1$ denotes the estimate of $\boldsymbol{\theta}_1^*$; for a sufficiently large N, there exists a radius r_1 such that $\tilde{\boldsymbol{\theta}}_1$ is uniquely defined, with ξ-probability 1. Hence, by restricting to Θ_1, the parameter $\boldsymbol{\theta}_1^*$ is identifiable.

The properties of $\hat{\bar{Y}}_{nn}^{\mathrm{mc}}$ are stated in the following theorem, the proof of which relies on some technical lemmas collected in the Appendix.

Theorem 1. Assume that (a)–(g) hold. Partition the parameter space as in Remark 3 and restrict to Θ_1, say. Then we have the following results:

1. Design consistency. $\hat{\bar{Y}}_{nn}^{\mathrm{mc}}$ is consistent for $\bar{Y}$ in the sense that $\lim_{\nu\to\infty} P(|\hat{\bar{Y}}_{nn}^{\mathrm{mc}} - \bar{Y}| < \epsilon) = 1$ with ξ-probability 1 and for any fixed $\epsilon > 0$.

2. Asymptotic normality. The asymptotic distribution of $\hat{\bar{Y}}_{nn}^{\mathrm{mc}}$ is such that

$$\frac{\hat{\bar{Y}}_{nn}^{\mathrm{mc}} - \bar{Y}}{\sqrt{V(\tilde{\bar{Y}}_{nn}^{\mathrm{mc}})}} \to \mathrm{N}(0, 1), \tag{16}$$

where

$$V(\tilde{\bar{Y}}_{nn}^{\mathrm{mc}}) = \frac{1}{N^2} \sum_i^N \sum_j^N (\pi_{ij} - \pi_i \pi_j) \frac{(y_i - \tilde{f}_i \tilde{\boldsymbol{\beta}}_{nn})}{\pi_i} \frac{(y_j - \tilde{f}_j \tilde{\boldsymbol{\beta}}_{nn})}{\pi_j}, \tag{17}$$

$\tilde{\bar{Y}}_{nn}^{\mathrm{mc}}$ is the generalized difference estimator

$$\tilde{\bar{Y}}_{nn}^{\mathrm{mc}} = \hat{\bar{Y}} + \left\{ N^{-1} \sum_{i=1}^N \tilde{f}_i - N^{-1} \sum_{i=1}^n d_i \tilde{f}_i \right\} \tilde{\boldsymbol{\beta}}_{nn}, \tag{18}$$

$$\tilde{\boldsymbol{\beta}}_{nn} = \frac{\sum_{i=1}^N q_i (\tilde{f}_i - \bar{f})(y_i - \bar{Y})}{\sum_{i=1}^N q_i (\tilde{f}_i - \bar{f})^2}, \tag{19}$$

and $\bar{f} = N^{-1} \sum_{i=1}^N \tilde{f}_i$.

For the proof, see the Appendix.

The next result shows that the variance of the asymptotic distribution of $\hat{\bar{Y}}_{nn}^{\mathrm{mc}}$ can be estimated consistently under mild assumptions. This result also holds for the estimator of the variance of the ordinary model calibration estimators proposed by Wu and Sitter (2001, sec. 3.2) for fixed-size sampling designs.

Theorem 2. Assume that (a)–(g) hold. Partition the parameter space as in Remark 3 and restrict to Θ_1, say. Then

$$v(\hat{\bar{Y}}^{\mathrm{mc}}_{nn}) = \frac{1}{N^2}\sum_i^n\sum_j^n \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{e_i}{\pi_i}\frac{e_j}{\pi_j}, \tag{20}$$

where $e_i = y_i - \hat{f}_i\hat{\beta}_{nn}$, is design consistent for $V(\tilde{\bar{Y}}^{\mathrm{mc}}_{nn})$.

For the proof see the Appendix.

In the following corollary, the pivotal considered in (16) is modified to account for this result.

Corollary 1. Assume that (a)–(g) hold. Partition the parameter space as in Remark 3 and restrict to Θ_1, say. Then, as $\nu \to \infty$, $(\hat{\bar{Y}}^{\mathrm{mc}}_{nn} - \bar{Y})/\sqrt{v(\hat{\bar{Y}}^{\mathrm{mc}}_{nn})} \to \mathrm{N}(0,1)$, where $v(\hat{\bar{Y}}^{\mathrm{mc}}_{nn})$ is given in (20).

Proof. The result follows from Theorem 2, for which $v(\hat{\bar{Y}}^{\mathrm{mc}}_{nn})/V(\tilde{\bar{Y}}^{\mathrm{mc}}_{nn})$ converges in probability to 1.

5. A LOCAL POLYNOMIAL MODEL CALIBRATION ESTIMATOR

In this section nonparametric model calibration is performed by means of local polynomials. Here we add in the definition of the local polynomial regression estimator in (6) a regression step to gain the property of calibration with respect to the working model. The steps to define this estimator mirror the ones used to obtain $\hat{\bar{Y}}^{\mathrm{mc}}_{nn}$, whereas the methodological assumptions are taken from the derivation of $\hat{\bar{Y}}_{lp}$. As a consequence, we consider a single auxiliary variable to make the theory more tractable. As for $\hat{\bar{Y}}_{lp}$, the properties of the estimator are expected to hold for multivariate $\mathbf{x}_i$; however, the curse of dimensionality would make practical applications complicated. Generalized additive models (Opsomer et al. 2001, 2003) can then be considered.

We assume that the finite population of the y_i's conditioned on the x_i's is a realization from a superpopulation ξ described by model (4). The *local polynomial model calibration estimator*, $\hat{\bar{Y}}^{\mathrm{mc}}_{lp} = N^{-1}\sum_{i=1}^n w_i y_i$, is obtained by seeking weights w_i that minimize the distance measure Φ_s in (1), under the constraints $N^{-1}\sum_{i=1}^n w_i = 1$ and $N^{-1}\sum_{i=1}^n w_i\hat{m}_i = N^{-1}\sum_{i=1}^N \hat{m}_i$, where fitted values $\hat{m}_i$ are obtained by means of (5). Minimization problem is solved as for $\hat{\bar{Y}}^{\mathrm{mc}}_{nn}$ and provides as the resulting estimator

$$\hat{\bar{Y}}^{\mathrm{mc}}_{lp} = \hat{\bar{Y}} + \frac{1}{N}\left\{\sum_{i=1}^N \hat{m}_i - \sum_{i=1}^n d_i\hat{m}_i\right\}\hat{\beta}_{lp}, \tag{21}$$

where $\hat{\beta}_{lp}$ takes the same form as $\hat{\beta}_{nn}$ in (14) but with fitted values $\hat{m}_i$ obtained by means of local polynomial smoothing instead of neural networks. With respect to $\hat{\bar{Y}}_{lp}$ in (6), the added calibration step performed by $\hat{\beta}_{lp}$ will asymptotically make $\hat{\bar{Y}}^{\mathrm{mc}}_{lp}$ more efficient than $\hat{\bar{Y}}_{lp}$ when population fits for y are biased, as we noted in the discussion after (13). In the context of local polynomials, this might happen for local constant fits with larger bandwidths.

Theorem 3 states that $\hat{\bar{Y}}^{\mathrm{mc}}_{lp}$ is asymptotically design unbiased and consistent. Moreover, its asymptotic distribution is derived from that of the generalized difference-type estimator

$$\tilde{\bar{Y}}^{\mathrm{mc}}_{lp} = \hat{\bar{Y}} + \left\{N^{-1}\sum_{i=1}^N \tilde{m}_i - N^{-1}\sum_{i=1}^n d_i\tilde{m}_i\right\}\tilde{\beta}_{lp}, \tag{22}$$

where $\tilde{m}_i$, for $i = 1,\ldots,N$, are the fitted values at the population level defined as

$$\tilde{m}_i = \mathbf{e}_1'(\mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{W}_i\mathbf{y}. \tag{23}$$

In the latter, $\mathbf{y} = (y_1,\ldots,y_N)'$, $\mathbf{W}_i = \mathrm{diag}\{K_h(x_j - x_i)\}_{j\in\mathcal{U}}$, and $\mathbf{X}_i = [1 \;\; x_j - x_i \;\; \cdots \;\; (x_j - x_i)^p]_{j\in\mathcal{U}}$; further, $\tilde{\beta}_{lp} = \sum_{i=1}^N q_i(\tilde{m}_i - \bar{m})(y_i - \bar{Y})/\sum_{i=1}^N q_i(\tilde{m}_i - \bar{m})^2$ and $\bar{m} = N^{-1}\times\sum_{i=1}^N \tilde{m}_i$. The asymptotic framework is as in Section 4, whereas the regularity conditions assumed are those considered by Breidt and Opsomer (2000, sec. 1.3) and reported in the Appendix. The value of the bandwidth h is considered fixed here as for $\hat{\bar{Y}}_{lp}$; a discussion on how to select values for h is deferred to Section 6.4.

Theorem 3. Assume (A1)–(A7) in the Appendix. Then we have the following results:

1. Asymptotic design unbiasedness and consistency. The local polynomial model calibration estimator $\hat{\bar{Y}}^{\mathrm{mc}}_{lp}$ is asymptotically design unbiased in the sense that $\lim_{\nu\to\infty} E(\hat{\bar{Y}}^{\mathrm{mc}}_{lp} - \bar{Y}) = 0$ with ξ-probability 1, and is design-consistent in the sense that $\lim_{\nu\to\infty} P(|\hat{\bar{Y}}^{\mathrm{mc}}_{lp} - \bar{Y}| < \epsilon) = 1$ with ξ-probability 1 and for any fixed $\epsilon > 0$.
2. Asymptotic normality. $(\tilde{\bar{Y}}^{\mathrm{mc}}_{lp} - \bar{Y})/\sqrt{V(\tilde{\bar{Y}}^{\mathrm{mc}}_{lp})} \to \mathrm{N}(0,1)$, as $\nu\to\infty$, where

$$V(\tilde{\bar{Y}}^{\mathrm{mc}}_{lp}) = \frac{1}{N^2}\sum_i^N\sum_j^N(\pi_{ij} - \pi_i\pi_j)\frac{R_i}{\pi_i}\frac{R_j}{\pi_j}, \tag{24}$$

with $R_i = y_i - \tilde{m}_i\tilde{\beta}_{lp}$, implies that

$$(\hat{\bar{Y}}^{\mathrm{mc}}_{lp} - \bar{Y})\Big/\sqrt{V(\tilde{\bar{Y}}^{\mathrm{mc}}_{lp})} \to \mathrm{N}(0,1). \tag{25}$$

For the proof see the Appendix.

As for $\hat{\bar{Y}}^{\mathrm{mc}}_{nn}$, we now introduce a design-consistent estimator of the variance of the asymptotic distribution of $\hat{\bar{Y}}^{\mathrm{mc}}_{lp}$, and consequently modify the pivotal in (25) to account for this result.

Theorem 4. Assume that (A1)–(A7) in the Appendix hold. Then

$$v(\hat{\bar{Y}}^{\mathrm{mc}}_{lp}) = \frac{1}{N^2}\sum_i^n\sum_j^n \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{r_i}{\pi_i}\frac{r_j}{\pi_j}, \tag{26}$$

where $r_i = y_i - \hat{m}_i\hat{\beta}_{lp}$, is design-consistent for $V(\tilde{\bar{Y}}^{\mathrm{mc}}_{lp})$.

For the proof, see the Appendix.

Corollary 2. Assume that (A1)–(A7) in the Appendix hold. Then, as $\nu\to\infty$, $(\hat{\bar{Y}}^{\mathrm{mc}}_{lp} - \bar{Y})/\sqrt{v(\hat{\bar{Y}}^{\mathrm{mc}}_{lp})} \to \mathrm{N}(0,1)$.

Proof. The result follows from Theorem 4, for which $v(\hat{\bar{Y}}^{\mathrm{mc}}_{lp})/V(\tilde{\bar{Y}}^{\mathrm{mc}}_{lp})$ converges in probability to 1.

6. SIMULATION STUDIES

In this section we report on some simulation experiments carried on to investigate the finite-sample performance of the proposed estimators of $\bar{Y}$. To allow comparisons, the design and structure of this investigation is taken from the simulation study conducted by Breidt and Opsomer (2000), where a single auxiliary variable is considered. Nevertheless, some features have also been changed and introduced to provide new insights into the topic. The simulation studies compare the behavior of the following estimators of $\bar{Y}$:

$\hat{\bar{Y}}$,	Horvitz–Thompson
$\hat{\bar{Y}}_c$,	Calibration–linear regression [see (2)]
$\hat{\bar{Y}}_{nn}^{\mathrm{mc}}$,	Neural network model calibration [see (13)]
$\hat{\bar{Y}}_{lp}$,	Local polynomial regression [see (6)]
$\hat{\bar{Y}}_{lp}^{\mathrm{mc}}$,	Local polynomial model calibration [see (21)].

The first two estimators are parametric estimators, in that they assume a constant and a linear model for the regression function of the survey variable. The other estimators allow for more complex modeling of the regression function. Nonparametric estimators can be considered classes of estimators, because they all depend on the values taken by different model selection parameters. Namely, local polynomial estimators depend on the order of the local expansion, on the choice of the kernel function K, and of values taken by the bandwidth h. In contrast, neural network estimators depend on the number of units in the hidden layer M and the weight decay parameter λ, as shown in (7) and (12). Because these parameters range over their allowed values, different estimators of the mean are generated.

We consider model selection in a presampling perspective; that is, the values of model selection parameters are determined in advance and kept fixed in repeated sampling. Following Breidt and Opsomer (2000), for local polynomial estimators, local constant and local linear estimators have been considered; the Epanechninkov kernel, defined as $K(t) = .75(1 - t^2)$ if $|t| < 1$ and $K(t) = 0$ otherwise, has been used for all kernel based estimators with the same two different bandwidth values, $h = .1$ and $h = .25$. The same values have been considered to compare results. Higher-order polynomials, such as local quadratic or local cubic approximations, have not been considered; although they provide a better approximation for internal points, they pay the price of far more erratic behavior on the boundaries and in presence of extreme values.

Along with bandwidth selection for local polynomials, the choice of complexity parameters for neural networks has always been a challenging issue. To better understand the behavior of neural networks in this particular setting of model calibration, the complexity parameters have been kept fixed over repeated sampling and chosen to have quite a wide range of possible scenarios; that is, we allowed λ and M to take different combination of values, to investigate their influence on the efficiency of the resulting estimator. In the present work we report on only five of them, to make reporting more tractable; detailed results are available on request. We report on estimators calculated by setting $M = 2$ and $\lambda = 25\mathrm{e}{-4}$, $M = 3$ and $\lambda = 5\mathrm{e}{-3}$, $M = 4$ and $\lambda = 5\mathrm{e}{-3}$, $M = 6$ and $\lambda = 1\mathrm{e}{-2}$, and $M = 8$ and $\lambda = 1\mathrm{e}{-2}$ (see Ripley 1996, sec. 5.5, for rules on the choice of λ). Values of $M > 8$ provide estimators whose performance is virtually the same as when $M = 8$ is used, with λ kept constant. Values of $\lambda > 1\mathrm{e}{-2}$ for these nets provide the same results as a small value for both M and λ, and thus these results are not reported here. Neural networks have actually been fitted by means of the R function nnet( ), which uses a quasi-Newton optimizer, but other free and commercial software packages are available. Values of the inputs and the output should be scaled to the range [0, 1] to aid convergence of the optimizer. The activation function has been chosen to be logistic.

Survey variables have been generated according to eight different models. Each of these models is characterized by a univariate regression function, or signal, that is, $E_\xi(y_k|x) = f_k(x)$ for $k = 1, \ldots, 8$ and $x \in \mathcal{R}$. We considered the following regression functions:

Linear:	$f_1(x) = 1 + 2(x - .5)$
Quadratic:	$f_2(x) = 1 + 2(x - .5)^2$
Bump:	$f_3(x) = 1 + 2(x - .5) + \exp(-200(x - .5)^2)$
Jump:	$f_4(x) = 1 + 2(x - .5)\mathbb{I}(x \le .65) + .65\mathbb{I}(x > .65)$
cdf:	$f_5(x) = \Phi((.5 - 2x)/.02)$, where Φ is the standard normal cdf
Exponential:	$f_6(x) = \exp(-8x)$
Cycle1:	$f_7(x) = 2 + \sin(2\pi x)$
Cycle4:	$f_8(x) = 2 + \sin(8\pi x)$,

with $x \in [0, 1]$. Breidt and Opsomer (2000) described the choice of such signals when the population values for x are generated as iid uniform on [0, 1] random variables. We considered this scenario and a skewed distribution for x as well; that is, we also conducted simulations for which the auxiliary variable is iid from a beta distribution with expected value 2/7 and variance 7/196.

Population values for all survey variables but the fifth one have been generated from the regression functions by adding mean-0 normal errors with variance such that the signal-to-noise ratio would approximately equal 4:1 for all populations. This implies that approximately 20% of the variance of the survey variables is due to the error. The cdf population, in contrast, consists of binary measurements generated from the linear population as $y_{5i} = \mathbb{I}(y_{1i} \le .5)$. Hence the finite population mean of y_5 is the population cdf of y_1 at the point $t = .5$. Using the same estimation strategy for continuous survey variables and for a binary one could be debatable. Even though more suitable networks can be chosen to account for a binary response, here we use the same one for all populations, to allow comparisons. The effective value of the proportion of variance due to noise is defined as

$$\mathrm{VP} = (S_y^2 - S_f^2)/S_y^2, \tag{27}$$

where S_y^2 is the population variance of the survey variable and S_f^2 is the population variance of the corresponding signal. For each simulation, 1,000 samples of size $n = 100$ have been drawn by simple random sampling from a population of size $N = 1{,}000$ and the estimators calculated together with their variance estimators. Given an estimator $\hat{\bar{Y}}_*$, its performance is evaluated using the following quantities:

- Relative bias: $RelB(\hat{\bar{Y}}_*) = (\widehat{E}(\hat{\bar{Y}}_*) - \bar{Y})/\bar{Y}$, where $\widehat{E}$ denotes the Monte Carlo estimate of the expected value.

- Scaled mean squared error (SMSE), defined as

$$SMSE(\hat{\bar{Y}}_*) = \frac{\widehat{MSE}(\hat{\bar{Y}}_*)}{\widehat{MSE}(\bar{y})\text{VP}}, \tag{28}$$

  where $\widehat{MSE}$ is the Monte Carlo estimate of the MSE and $\bar{y}$ is the sample mean, that is, the Horvitz–Thompson estimator for this design.
- Relative bias of a variance estimator: $RelB(v(\hat{\bar{Y}}_*)) = [\widehat{E}(v(\hat{\bar{Y}}_*)) - \widehat{MSE}(\hat{\bar{Y}}_*)]/\widehat{MSE}(\hat{\bar{Y}}_*)$.

Note that with $SMSE(\hat{\bar{Y}}_*)$, we compare the MSE of an estimator with its lowest possible value. In fact, the MSE of the sample mean times the proportion of variance of y due to noise can be considered the MSE of an ideal estimator that perfectly captures the signal, and whose left variation is due only to the irreducible error of the noise. Hence the smaller the value taken by SMSE, the greater the efficiency of the estimator. We first report on the study dealing with the auxiliary variable x generated from a uniform distribution, and then move on to the one based on a skewed variable x generated from a beta distribution.

6.1 Simulation With a Uniform x

Results for the SMSE of the estimators in this simulation are reported in Table 1. The first row of the table gives the values taken by VP for each population. Attention is focused on the behavior of the class of nonparametric estimators rather than on a single estimator; that is, we are interested in the efficiency of the class of estimators, irrespective of the choice of the complexity parameters. It is well known that nonparametric methods are usually sensible to the choice of such parameters, different values of which may lead to very different fitted values. Because model selection may not be feasibly conducted for all survey variables, when more than one variable is of concern, understanding the behavior of the estimators for a range of values of the complexity parameters is of greater interest.

First, we note that neural network estimators all behave similarly with respect to the choice of the number of units in the hidden layer. In contrast, estimators based on local polynomials are more erratic in correspondence of different values of the bandwidth and of the order of the local fit. Overall, nonparametric estimators lead to good gains in efficiency with respect to the regression estimator in all populations but the linear one. The SMSE values for the best nonparametric model calibration estimators are always extremely close to 1 for most populations. For the Cycle4 population, the regression function is extremely complex (a sinusoid completing four full cycles on [0, 1]), and performance of the nonparametric estimators varies widely. The more the complexity parameters allow approximation of more complex functions, the greater the gain in efficiency. For neural networks, this is clearly shown by the decrease in SMSE with increasing M. The same is true for local polynomials with a smaller bandwidth.

Last two columns in Table 1 give an indication when robustness over different populations is of interest. They report the average value of SMSE over all populations (Ave8) and after removing the last population (Ave7). First, we note that neural network estimators are almost always more efficient than the others, particularly when M is high. This is due to the fact that most of the fitted values averaged over repeated sampling obtained with neural networks are indistinguishable from the real mean function. Differences are seen only for the last population. Averaged fitted values are usually less well behaved for local polynomials, and even with a small value of the bandwidth, they cannot completely capture the patterns of the last population; further details on this are available on request.

Local polynomial model calibration estimators are on average slightly more efficient than the corresponding local polynomial regression estimators, with substantial improvements when fitted values are obtained with the large bandwidth and a local constant fit. In this case, much of the error of the fits in

Table 1. SMSE of the Investigated Estimators for the Eight Populations and Averages Over the First Seven and All of the Eight Populations—Uniform x

	Linear	*Quad*	*Bump*	*Jump*	*cdf*	*Exp*	*Cycle1*	*Cycle4*	*Ave7*	*Ave8*
VP	.201	.200	.218	.203	.327	.190	.178	.294		
$\bar{y}$	4.970	5.000	4.583	4.933	3.060	5.258	5.603	3.396	4.772	4.600
$\hat{\bar{Y}}_c$	1.029	5.177	1.602	1.196	1.208	3.151	2.896	3.320	2.323	2.448
$\hat{\bar{Y}}_{nn}^{\text{mc}}$										
$M=2$, $\lambda=25\text{e}{-4}$	1.085	1.024	1.104	1.089	1.146	1.126	1.086	2.741	1.094	1.300
$M=3$, $\lambda=5\text{e}{-3}$	1.074	1.018	1.167	1.089	1.140	1.127	1.084	2.129	1.100	1.229
$M=4$, $\lambda=5\text{e}{-3}$	1.075	1.016	1.136	1.094	1.138	1.127	1.085	1.315	1.096	1.123
$M=6$, $\lambda=1\text{e}{-2}$	1.060	1.016	1.244	1.086	1.123	1.133	1.080	.943	1.106	1.086
$M=8$, $\lambda=1\text{e}{-2}$	1.062	1.014	1.229	1.083	1.127	1.129	1.080	.898	1.103	1.078
$\hat{\bar{Y}}_{lp}^{\text{mc}}$										
$p=0$, $h=.1$	1.085	1.066	1.078	1.077	1.138	1.236	1.114	1.233	1.113	1.128
$p=0$, $h=.25$	1.072	1.234	1.322	1.090	1.110	1.633	1.421	3.383	1.269	1.533
$p=1$, $h=.1$	1.191	1.050	1.114	1.115	1.145	1.204	1.164	1.238	1.140	1.153
$p=1$, $h=.25$	1.070	1.023	1.352	1.064	1.102	1.215	1.116	3.287	1.134	1.404
$\hat{\bar{Y}}_{lp}$										
$p=0$, $h=.1$	1.080	1.084	1.075	1.077	1.133	1.265	1.122	1.496	1.119	1.167
$p=0$, $h=.25$	1.092	1.631	1.342	1.118	1.106	1.889	1.476	3.619	1.379	1.659
$p=1$, $h=.1$	1.193	1.049	1.107	1.113	1.139	1.207	1.172	1.427	1.140	1.176
$p=1$, $h=.25$	1.072	1.038	1.350	1.070	1.094	1.216	1.318	3.358	1.166	1.440

NOTE: The first row displays the proportion of the variance of the survey variables due to noise.

estimating the regression function is due to bias: $\hat{\beta}_{lp}$ takes value on average substantially different from 1, and so the additional calibration step provides some improvement in efficiency. This is not so in the other cases.

Values of *RelB* are less than 1% for all estimators and thus are not presented here. In contrast, variance estimators underestimate the Monte Carlo MSE in most cases, especially when the nonparametric method overfits the data. In fact, because the estimators of the variance are all based on residuals from the fitted values, the harder the nonparametric method fits the data, the smaller the residuals (and hence the variance estimator), the smaller the generalization power, and hence the larger the MSE. The relative bias ranges from 6% to 23% for most populations, with the exception of the Cycle4 population, for which the relative bias increases up to 30%. Sample sizes larger than the one considered here are likely required, to reduce underestimation of the variance estimator.

6.2 Simulation With a Skewed *x*

Population values for the auxiliary variable in this simulation have been generated from the beta distribution introduced earlier. Results for the SMSE of the estimators, together with the values taken by VP for each population, are reported in Table 2. Relative biases of all estimators are again negligible.

In this case, differences are more striking. Neural network estimators perform rather well across all populations, with a small degree of variability over the different values of M and λ. Their SMSE takes values close to 1 for most populations. Exceptions are observed only for the Cycle4 population. In contrast, the efficiency of local polynomial estimators varies widely across populations and shows large losses in efficiency in several cases. This is particularly true when we calculate fitted values by means of a local linear fit and the small bandwidth. The performance of $\hat{\bar{Y}}_{lp}$ and $\hat{\bar{Y}}^{\mathrm{mc}}_{lp}$ in this case is quite poor for all populations. This may be explained by the fact that with this positively skewed distribution there is a boundary of the support of x, and consequently of the survey variables, which is less densely populated. Hence there are samples for which points on the boundary do not provide a reasonable local approximation when the bandwidth is too small. The additional regression step performed by $\hat{\bar{Y}}^{\mathrm{mc}}_{lp}$ with respect to $\hat{\bar{Y}}_{lp}$ again provides some improvement in efficiency. Variance estimators are still negatively biased in all populations but the linear one. Relative bias usually takes values ranging from 5% to 24%, with a large value for the Cycle4 population of about 43%. This relatively poor performance suggests that further investigation on variance estimation is needed for particularly complex regression functions.

6.3 Real Data Application

The mid-Atlantic highlands region includes the area from the Blue Ridge Mountains in the east to the Ohio River in the west and from the Catskill Mountains in the north to North Carolina in the south. In the years 1993–1996 more than 500 stream reaches across this region were sampled and visited, and some revisited, to assess their condition in terms of the chemistry and health of the biological organisms (EPA 2000). Among the factors affecting stream condition, high concentrations of nitrogen and phosphorus are symptoms of excessive nutrients introduced into the stream. This phenomenon would likely increase algal growth, thereby depleting the oxygen in the water and choking out other forms of biota and significantly altering the animal communities present. One possible cause of nutrient enrichment in streams is agricultural fertilizer application to fields. The proportion of land devoted to agriculture in a particular watershed can be obtained from remote sensing and thus is available for all stream locations without the need to go onsite. A square-root transformation of this independent variable overcomes the problem of concentration of points on small values. Figure 2 gives scatterplots of total nitrogen (NTL) and total phosphorus (PTL) concentrations with respect to the square root of the proportion of agricultural land (AG) for 574 streams.

Table 2. SMSE of the Investigated Estimators for the Eight Populations and Averages Over the First Seven and All of the Eight Populations—Skewed x

	Linear	*Quad*	*Bump*	*Jump*	*cdf*	*Exp*	*Cycle1*	*Cycle4*	*Ave7*	*Ave8*
VP	.202	.230	.208	.197	.385	.172	.182	.197		
$\bar{y}$	4.949	4.347	4.797	5.078	2.598	5.805	5.481	5.084	4.722	4.767
$\hat{\bar{Y}}_c$	1.002	1.930	1.580	.963	1.223	2.758	4.030	4.981	1.927	2.308
$\hat{\bar{Y}}^{\mathrm{mc}}_{nn}$										
$M=2$, $\lambda=25\mathrm{e}{-4}$	1.024	.918	1.193	1.003	1.058	1.264	1.125	4.186	1.084	1.471
$M=3$, $\lambda=5\mathrm{e}{-3}$	1.019	.914	1.222	.987	1.052	1.268	1.115	3.400	1.082	1.372
$M=4$, $\lambda=5\mathrm{e}{-3}$	1.024	.915	1.205	.997	1.056	1.262	1.116	2.195	1.082	1.221
$M=6$, $\lambda=1\mathrm{e}{-2}$	1.010	.911	1.245	.984	1.046	1.261	1.111	1.616	1.081	1.148
$M=8$, $\lambda=1\mathrm{e}{-2}$	1.009	.912	1.233	.988	1.045	1.260	1.111	1.535	1.080	1.137
$\hat{\bar{Y}}^{\mathrm{mc}}_{lp}$										
$p=0$, $h=.1$	1.148	.990	1.088	1.173	1.049	1.364	1.137	2.040	1.136	1.249
$p=0$, $h=.25$	1.073	1.264	1.419	1.040	1.150	1.911	1.626	5.018	1.355	1.813
$p=1$, $h=.1$	1.629	1.404	1.585	2.574	1.040	3.471	1.883	2.023	1.941	1.951
$p=1$, $h=.25$	1.033	.937	1.317	1.168	1.066	1.379	1.306	4.948	1.172	1.644
$\hat{\bar{Y}}_{lp}$										
$p=0$, $h=.1$	1.157	1.020	1.107	1.192	1.049	1.446	1.171	2.394	1.163	1.317
$p=0$, $h=.25$	1.336	1.652	1.605	1.387	1.220	2.527	2.025	5.420	1.679	2.146
$p=1$, $h=.1$	1.631	1.403	1.557	2.610	1.038	3.530	1.849	2.274	1.945	1.986
$p=1$, $h=.25$	1.036	.953	1.338	1.171	1.069	1.384	1.339	5.065	1.184	1.669

NOTE: The first row displays the proportion of the variance of the survey variables due to noise.

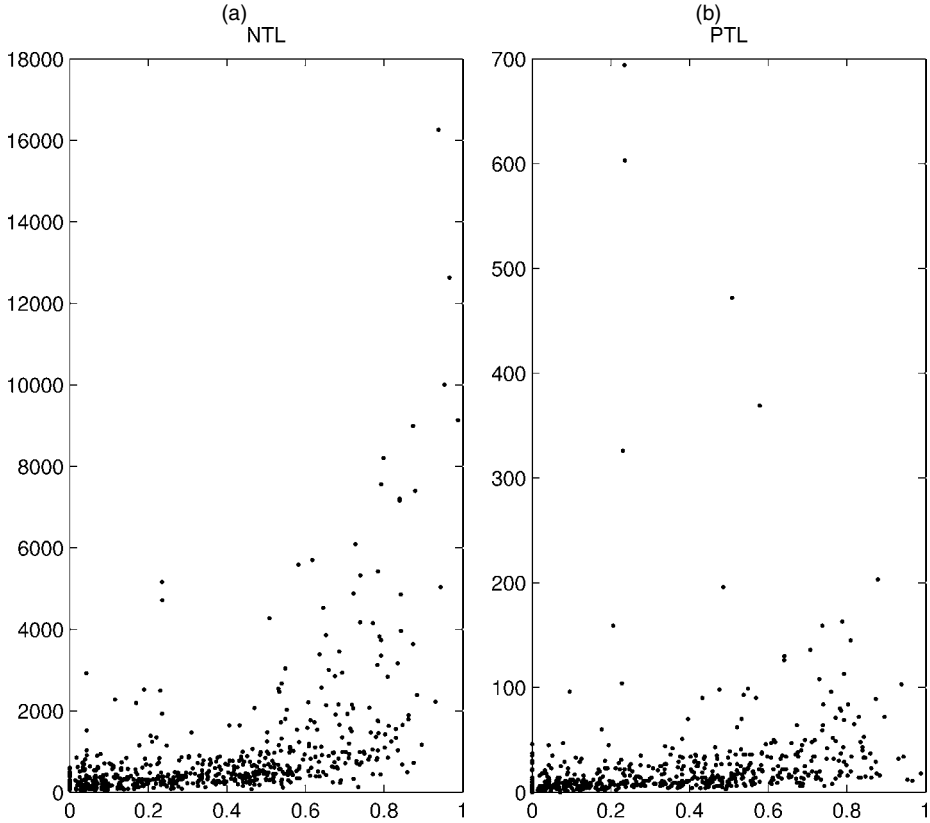

Figure 2. Scatterplots for the (a) NTL and (b) PTL Concentrations, With Respect to the Square Root of the Proportion of Agricultural Land.

Only the first visit for each stream has been considered. Despite the presence of numerous influential observations, a linear regression model would seem adequate to determine the relationship of PTL and AG. In contrast, a more complex structure to determine the relationship of NTL and AG might be considered. To investigate whether nonparametric model calibration could be of use in such a context, we conducted a simulation study. In particular, we considered the set of $N = 574$ streams as a finite population for which NTL and PTL are survey variables of interest. Moreover, AG can be considered an auxiliary variable whose value is available for each unit in the population from remote sensing. For each survey variable, we selected 5,000 random samples without replacement of $n = 100$ units. For each sample, we calculated and evaluated the performance of the same set of estimators considered in the previous section. Because the true mean function in the relationship between the survey variables and the auxiliary variable is unknown, the relative efficiency of the estimators is defined as

$$Eff(\hat{\bar{Y}}_*) = \widehat{MSE}(\hat{\bar{Y}}_*) \big/ \widehat{MSE}(\hat{\bar{Y}}_c). \tag{29}$$

Table 3 gives the values of the efficiency for both survey variables. Good gains in efficiency with respect to the calibration estimator are provided for NTL by all neural network estimators almost independent of the choice of the complexity parameters. Moreover, negligible losses in efficiency are shown for PTL. In contrast, estimators based on local polynomials provide almost the same good performance as that of $\hat{\bar{Y}}^{\mathrm{mc}}_{nn}$ in all cases but one. When fitted values are obtained through a local linear fit with a small bandwidth, the performance of the resulting estimators is really poor. This behavior can be explained by the presence of extreme points that, when sampled and considered in a local linear fit with few observations, provide unreasonable approximations. $\hat{\bar{Y}}^{\mathrm{mc}}_{lp}$ does not perform much better than $\hat{\bar{Y}}_{lp}$, because inefficiency in this situation is a consequence of overfitting the data. This problem is overcome by kernel approximations by means of a more robust local constant fit.

Table 3. Efficiency of the Investigated Estimators for the NTL and PTL Survey Variables With Respect to the Calibration Estimator [eq. (29)]

	NTL	*PTL*
$\bar{y}$	1.289	1.033
$\hat{\bar{Y}}_c$	1.000	1.000
$\hat{\bar{Y}}^{\mathrm{mc}}_{nn}$		
$M = 2$, $\lambda = 25\mathrm{e}{-4}$	.809	1.007
$M = 3$, $\lambda = 5\mathrm{e}{-3}$	.803	1.004
$M = 4$, $\lambda = 5\mathrm{e}{-3}$	.802	1.004
$M = 6$, $\lambda = 1\mathrm{e}{-2}$	.803	1.001
$M = 8$, $\lambda = 1\mathrm{e}{-2}$	.802	1.001
$\hat{\bar{Y}}^{\mathrm{mc}}_{lp}$		
$p = 0$, $h = .1$	.812	1.084
$p = 0$, $h = .25$	.812	1.021
$p = 1$, $h = .1$	9.390	1.928
$p = 1$, $h = .25$	.830	1.019
$\hat{\bar{Y}}_{lp}$		
$p = 0$, $h = .1$	.830	1.036
$p = 0$, $h = .25$	.845	1.011
$p = 1$, $h = .1$	9.610	1.976
$p = 1$, $h = .25$	.835	1.015

6.4 Model Selection for Nonparametric Model Calibration

In this work, complexity parameters for both $\hat{\bar{Y}}^{\mathrm{mc}}_{nn}$ and $\hat{\bar{Y}}^{\mathrm{mc}}_{lp}$ have been considered as fixed quantities. This means that for local polynomials, for example, the value of h used to compute population fits in (23) is the same one used to compute sample fits in (5). In other words, here (as in Breidt and Opsomer 2000) the issue of how to best select a value for h from sample data has not been addressed as that of selecting M and λ for neural networks.

While revising this article, we became aware of a work by Opsomer and Miller (2004) addressing this issue for the local polynomial regression estimator $\hat{\bar{Y}}_{lp}$. An optimal bandwidth value is defined to minimize the design MSE of $\hat{\bar{Y}}_{lp}$. This value does not have an explicit expression and thus is computed through simulation. A grid of values for h is considered, and the efficiency of the corresponding estimators is computed through repeated sampling. A cross-validation criterion modified to account for the design is then adopted to estimate this quantity from sample data. $\hat{\bar{Y}}_{lp}$ shows sample fits that can be written as a linear combination of y_i's, with weights that do not depend on the survey variable [see (5)]. This allows us to avoid computing such weights at each step of the leave-one-out procedure.

As described at the beginning of Section 6, we tried a wide range of values for M and λ and then kept them fixed over repeated sampling. This approach is the same as that followed by Opsomer and Miller (2004) to determine the optimal bandwidth value. However, fitted values used in $\hat{\bar{Y}}^{\mathrm{mc}}_{nn}$ are not linear combinations of the y_i's. This would require their computation at each leave-one-out step of the cross-validation procedure used to estimate such optimal values. This greatly increases the computational burden of the procedure. We are currently studying a modification of such cross-validation criterion for application to neural network nonparametric model calibration. We will report this elsewhere, along with an application of this procedure to $\hat{\bar{Y}}^{\mathrm{mc}}_{lp}$ and an investigation into the effects on its asymptotic properties.

7. CONCLUDING REMARKS

We have proposed and studied an application of nonparametric methods to the model calibration approach introduced by Wu and Sitter (2001) to the use of complete auxiliary information in complex surveys for estimating totals and means. Our application allows more flexible modeling by assuming more general models and uses nonparametric methods to obtain the fitted values on which to calibrate. We adopt neural network learning and local polynomial smoothing to estimate the functional relationship between the survey variable and the auxiliary variables. The resulting estimators are defined to account for the sampling design and are proven to be design-consistent. The moments of the asymptotic distribution are also derived, and a consistent estimator of the variance of each distribution is then proposed.

We investigated the performance of the proposed estimators for finite size samples through two simulation studies. We compared nonparametric model calibration estimators with nonparametric regression estimators and classical parametric ones and explore the effects of different distributions of the survey variables. Gains in efficiency with respect to the classical regression estimator are provided in all cases by neural network estimators, except when sampling from a linear population. An important pattern shown by neural networks is that once a weight decay parameter is included in the learning procedure, fitted values calculated using different numbers of units in the hidden layer provide estimators that display very similar behaviors. This is an interesting robustness result that puts less emphasis on model selection for neural networks; they exhibit very good performance even if the same structure is used for all populations. Different performance is exhibited only when approximating extremely complex functions.

In contrast, local polynomial estimators are much more sensible to the choice of the bandwidth value and the type of local approximation. Efficiency of the resulting estimators varies widely according to the selected values of the complexity parameters. The same structure may not be efficient enough for all survey variables, possibly leading to poor robustness. The local polynomial model calibration estimator has been shown to be more efficient than the corresponding local polynomial regression estimator when the regression function is estimated with bias. The calibration step performed by the model calibration estimator in this case recovers the efficiency lost by the approximating technique.

We have considered fixed values of M and λ for neural networks and of h for local polynomials over repeated sampling. The issue of theoretically optimal and practical selection of these values is of clear interest and optimality has to be defined in this context with respect to the design-based MSEs of the proposed estimators. Investigation in this regard will hopefully provide general guidelines. However, as noted previously, this issue seems to be less important for neural networks.

Simulation studies have been performed with one covariate. However, neural networks can easily accommodate multiple covariates, and we are currently investigating the performance of the proposed neural network estimator when applied to multivariate auxiliary information. This popular learning technique, although computer-intensive, works amazingly well in the presence of many records and many auxiliary variables. This would favor its application in survey organizations when auxiliary information at an individual level can be recovered; other circumstances than the one explored in this article through remote sensing also arise for survey organizations any time record linkage can be efficiently conducted between survey and census-level data. These situations are becoming more and more common, and the methods proposed here provide tools for using this type of complete information more efficiently.

APPENDIX: PROOFS AND REGULARITY CONDITIONS

Lemma A.1. Assume that (a)–(g) hold. Partition the parameter space as in Remark 3 and restrict to Θ_1, say. Then the design-based estimator of $\tilde{\boldsymbol{\theta}}$ obtained by (12) is such that $\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}} + O_p(n^{-1/2})$, where subscript 1 is dropped for ease of notation.

Proof. The proof is adapted from work of Wu (1999), who established this lemma for population parameters defined by estimating equations. First, $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ are weighted least squares estimates for a nonlinear function $f(\cdot)$. Existence of a solution to both (11) and (12) is then guaranteed by continuity of $f(\cdot)$ and compactness of the parameter space (Wu 1981). Restricting the parameter space to a subset

Θ_1 of Θ built as in Remark 3 provides uniqueness of both $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$. For ease of notation, we rewrite (11) and (12) as $\sum_{i=1}^{N} \boldsymbol{\zeta}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) = \mathbf{0}$ and $\sum_{i=1}^{n} d_i \boldsymbol{\zeta}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) = \mathbf{0}$. We can apply a Taylor series expansion to $N^{-1} \sum_{i=1}^{N} \boldsymbol{\zeta}(y_i, \mathbf{x}_i; \hat{\boldsymbol{\theta}})$ at $\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}$. We have

$$N^{-1} \sum_{i=1}^{N} \boldsymbol{\zeta}(y_i, \mathbf{x}_i; \hat{\boldsymbol{\theta}}) = N^{-1} \left\{ \sum_{i=1}^{N} \frac{\partial \boldsymbol{\zeta}(y_i, \mathbf{x}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \right\}' (\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) + o_p(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}). \quad \text{(A.1)}$$

By assumptions (c)–(e), and thus Remark 2, we have that

$$N^{-1} \sum_{i=1}^{n} d_i \boldsymbol{\zeta}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) = N^{-1} \sum_{i=1}^{N} \boldsymbol{\zeta}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) + O_p\big(n^{-1/2}\big). \quad \text{(A.2)}$$

Equation (A.2) calculated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ simplifies to $N^{-1} \sum_{i=1}^{N} \boldsymbol{\zeta}(y_i, \mathbf{x}_i; \hat{\boldsymbol{\theta}}) = O_p(n^{-1/2})$. The argument follows because by assumption (c), continuity of $f(\cdot)$ and compactness of the support of the $\mathbf{x}_i$'s and of the restricted parameter space, $N^{-1} \sum_{i=1}^{N} \partial \boldsymbol{\zeta}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = O(1)$.

Lemma A.2. Assume that (a)–(g) hold. Partition the parameter space as in Remark 3 and restrict to Θ_1, say. Then $N^{-1} \sum_{i=1}^{N} \hat{f}_i - N^{-1} \sum_{i=1}^{n} d_i \hat{f}_i = O_p(n^{-1/2})$.

Proof. Let us apply a Taylor series expansion to $\hat{f}_i = f(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$ at $\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}$. We obtain

$$f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) = f(\mathbf{x}_i, \tilde{\boldsymbol{\theta}}) + \left\{ \frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \right\}' (\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) + o_p(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}).$$

Now, by continuity of the function f and compactness of the support of the $\mathbf{x}_i$'s and of the restricted parameter space, we have that

$$\frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = O(1). \quad \text{(A.3)}$$

Hence, by Lemma A.1, we have

$$N^{-1} \sum_{i=1}^{N} \hat{f}_i = N^{-1} \sum_{i=1}^{N} \tilde{f}_i + O_p\big(n^{-1/2}\big) \quad \text{(A.4)}$$

and

$$N^{-1} \sum_{i=1}^{n} d_i \hat{f}_i = N^{-1} \sum_{i=1}^{n} d_i \tilde{f}_i + O_p\big(n^{-1/2}\big). \quad \text{(A.5)}$$

By assumptions (c)–(e), we also have $N^{-1} \sum_{i=1}^{N} \tilde{f}_i - N^{-1} \sum_{i=1}^{n} d_i \tilde{f}_i = O_p(n^{-1/2})$. This relation, together with (A.4) and (A.5), implies the argument.

Lemma A.3. Assume that (a)–(g) hold. Partition the parameter space as in Remark 3 and restrict to Θ_1, say. Then $N^{-1} \sum_{i=1}^{N} \hat{f}_i - N^{-1} \sum_{i=1}^{n} d_i \hat{f}_i = N^{-1} \sum_{i=1}^{N} \tilde{f}_i - N^{-1} \sum_{i=1}^{n} d_i \tilde{f}_i + O_p(n^{-1})$.

Proof. Consider a second-order Taylor series expansion of $f(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$ at $\hat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}$. Similarly to (A.3), we have that $\partial^2 f(\mathbf{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = O(1)$. This statement, together with (A.3) and Lemma A.1, implies that

$$N^{-1} \sum_{i=1}^{N} \hat{f}_i = N^{-1} \sum_{i=1}^{N} \tilde{f}_i + N^{-1} \left\{ \sum_{i=1}^{N} \frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \right\}' (\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) + O_p(n^{-1})$$

and

$$N^{-1} \sum_{i=1}^{n} d_i \hat{f}_i = N^{-1} \sum_{i=1}^{n} d_i \tilde{f}_i + N^{-1} \left\{ \sum_{i=1}^{n} d_i \frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \right\}' (\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) + O_p(n^{-1}).$$

By assumptions (c)–(e),

$$N^{-1} \left\{ \sum_{i=1}^{N} \frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \right\} - N^{-1} \left\{ \sum_{i=1}^{n} d_i \frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \right\} = O_p\big(n^{-1/2}\big).$$

Hence, by subtracting (A.5) from (A.4), the argument follows.

Lemma A.4. Assume that (a)–(g) hold. Partition the parameter space as in Remark 3 and restrict to Θ_1, say. Then $\hat{\beta}_{nn} = \tilde{\beta}_{nn} + O_p(n^{-1/2})$.

Proof. We can rewrite $\tilde{\beta}_{nn}$ as

$$\tilde{\beta}_{nn} = \frac{N^{-1} \sum_{i=1}^{N} q_i \tilde{f}_i y_i - N^{-1} \sum_{i=1}^{N} q_i \bar{f} \bar{Y}}{N^{-1} \sum_{i=1}^{N} q_i \tilde{f}_i^2 - N^{-1} \sum_{i=1}^{N} q_i \bar{f}^2}$$

and rewrite $\hat{\beta}_{nn}$ as

$$\hat{\beta}_{nn} = \frac{N^{-1} \sum_{i=1}^{n} d_i q_i \hat{f}_i y_i - N^{-1} \sum_{i=1}^{n} d_i q_i \breve{f} \breve{y}}{N^{-1} \sum_{i=1}^{n} d_i q_i \hat{f}_i^2 - N^{-1} \sum_{i=1}^{n} d_i q_i \breve{f}^2}. \quad \text{(A.6)}$$

Hence $\tilde{\beta}_{nn}$ can be seen as a function of population means. Each component of $\hat{\beta}_{nn}$ as written in (A.6) is a $\sqrt{n}$-consistent estimator of each of these means under the assumed regularity conditions, and the argument follows. Details are available on request.

Proof of Theorem 1

1. Design consistency. Let us consider the estimator introduced in (18). Being a generalized difference-type estimator, it is unbiased and consistent for $\bar{Y}$ for assumptions (c)–(e). Now, $\hat{\bar{Y}}_{nn}^{\mathrm{mc}}$ converges in probability to $\tilde{\bar{Y}}_{nn}^{\mathrm{mc}}$ because, by Lemmas A.3 and A.4, we can rewrite $\hat{\bar{Y}}_{nn}^{\mathrm{mc}}$ as

$$\begin{aligned} \hat{\bar{Y}}_{nn}^{\mathrm{mc}} &= \hat{\bar{Y}} + \left\{ N^{-1} \sum_{i=1}^{N} \tilde{f}_i - N^{-1} \sum_{i=1}^{n} d_i \tilde{f}_i \right\} \tilde{\beta}_{nn} + O_p(n^{-1}) \\ &= \tilde{\bar{Y}}_{nn}^{\mathrm{mc}} + O_p(n^{-1}). \end{aligned} \quad \text{(A.7)}$$

Therefore, $\hat{\bar{Y}}_{nn}^{\mathrm{mc}}$ converges in probability to $\bar{Y}$, and the argument follows.

2. Asymptotic normality. Convergence in probability implies convergence in distribution; therefore, $\hat{\bar{Y}}_{nn}^{\mathrm{mc}}$ inherits the limiting distribution of $\tilde{\bar{Y}}_{nn}^{\mathrm{mc}}$. A central limit theorem can be established for $\tilde{\bar{Y}}_{nn}^{\mathrm{mc}}$ from assumptions (c)–(e), and the result is established.

Proof of Theorem 2

Being the design variance of the Horvitz–Thompson estimator of the mean of the population residuals $y_i - \tilde{f}_i \tilde{\beta}_{nn}$, we have that $V(\tilde{\bar{Y}}_{nn}^{\mathrm{mc}}) = O(n^{-1})$. Hence it suffices to show that $v(\hat{\bar{Y}}_{nn}^{\mathrm{mc}}) - V(\tilde{\bar{Y}}_{nn}^{\mathrm{mc}}) = o_p(n^{-1})$. Let us consider the following estimator of $V(\tilde{\bar{Y}}_{nn}^{\mathrm{mc}})$:

$$v(\tilde{\bar{Y}}_{nn}^{\mathrm{mc}}) = \frac{1}{N^2} \sum_{i}^{n} \sum_{j}^{n} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i - \tilde{f}_i \tilde{\beta}_{nn}}{\pi_i} \frac{y_j - \tilde{f}_j \tilde{\beta}_{nn}}{\pi_j}.$$

From assumption (e), $v(\tilde{\bar{Y}}_{nn}^{\mathrm{mc}}) - V(\tilde{\bar{Y}}_{nn}^{\mathrm{mc}}) = o_p(n^{-1})$. Because $e_i e_j = (y_i - \tilde{f}_i \tilde{\beta}_{nn})(y_j - \tilde{f}_j \tilde{\beta}_{nn}) + O_p(n^{-1/2})$ by Lemma A.4 and $\hat{f}_i =$

$\tilde{f}_i + O_p(n^{-1/2})$ from Lemma A.2, $v(\hat{\bar{Y}}_{nn}^{\mathrm{mc}}) = v(\tilde{\bar{Y}}_{nn}^{\mathrm{mc}}) + o_p(n^{-3/2})$, and the argument follows.

Regularity Conditions for Theorem 3

(A1) For each ν, the x_i, for $i = 1, \ldots, N$, are iid $F(x) = \int_{-\infty}^{x} g(t)\,dt$, where $g(\cdot)$ is a density with compact support $[a_x, b_x]$ and $g(x) > 0$ for all $x \in [a_x, b_x]$.

(A2) For each ν, the x_i are considered fixed with respect to the superpopulation model ξ assumed in (4). The errors ε_i are independent and have mean 0, variance $v(x_i)$, and compact support, uniformly for each ν.

(A3) The mean function $m(\cdot)$ is continuous and has $p+1$ continuous derivatives, where p is the order of the local polynomial function. The variance function $v(x)$ is continuous and strictly positive.

(A4) The kernel $K(\cdot)$ has compact support $[-1, 1]$, is symmetric and continuous, and satisfies $\int_{-1}^{1} K(u)\,du = 1$.

(A5) As $\nu \to \infty$, $nN^{-1} \to \pi \in (0, 1)$, the bandwidth $h_\nu \to 0$ and $Nh_\nu^2/(\log\log N) \to \infty$.

(A6) For each ν, $\min_{i\in\mathcal{U}} \pi_i \geq \lambda > 0$, $\min_{i,j\in\mathcal{U}} \pi_{ij} \geq \lambda^* > 0$, and

$$\limsup_{\nu\to\infty} n \max_{i,j\in\mathcal{U}\,:\,i\neq j} |\pi_{ij} - \pi_i\pi_j| < \infty.$$

(A7) Additional assumptions involving higher-order inclusion probabilities are

$$\lim_{\nu\to\infty} n^2 \max_{(i_1,i_2,i_3,i_4)\in D_{4,N}} \left|E\left[\left(\delta_{i_1} - \pi_{i_1}\right)\left(\delta_{i_2} - \pi_{i_2}\right) \times \left(\delta_{i_3} - \pi_{i_3}\right)\left(\delta_{i_4} - \pi_{i_4}\right)\right]\right| < \infty,$$

where $D_{t,N}$ denotes the set of all distinct t-tuples $(i_1, i_2, \ldots, i_t)$ from $\mathcal{U}$,

$$\lim_{\nu\to\infty} \max_{(i_1,i_2,i_3,i_4)\in D_{4,N}} \left|E\left[\left(\delta_{i_1}\delta_{i_2} - \pi_{i_1 i_2}\right)\left(\delta_{i_3}\delta_{i_4} - \pi_{i_3 i_4}\right)\right]\right| = 0,$$

and

$$\limsup_{\nu\to\infty} n \max_{(i_1,i_2,i_3)\in D_{3,N}} \left|E\left[\left(\delta_{i_1} - \pi_{i_1}\right)^2\left(\delta_{i_2} - \pi_{i_2}\right)\left(\delta_{i_3} - \pi_{i_3}\right)\right]\right| < \infty.$$

Proof of Theorem 3

1. Asymptotic design unbiasedness and consistency. By the Markov inequality, it suffices to show that $\lim_{\nu\to\infty} E|\hat{\bar{Y}}_{lp}^{\mathrm{mc}} - \bar{Y}| = 0$. Write

$$\hat{\bar{Y}}_{lp}^{\mathrm{mc}} - \bar{Y} = \sum_{i=1}^{N} \frac{y_i - \tilde{m}_i\tilde{\beta}_{lp}}{N}\left(\frac{\delta_i}{\pi_i} - 1\right) + \sum_{i=1}^{N} \frac{\tilde{m}_i\tilde{\beta}_{lp} - \hat{m}_i\hat{\beta}_{lp}}{N}\left(\frac{\delta_i}{\pi_i} - 1\right). \quad \text{(A.8)}$$

By rewriting the right side of (A.8), we have that

$$\begin{aligned} E|\hat{\bar{Y}}_{lp}^{\mathrm{mc}} - \bar{Y}| &\leq E\left|\sum_{i=1}^{N} \frac{y_i - \tilde{m}_i\tilde{\beta}_{lp}}{N}\left(\frac{\delta_i}{\pi_i} - 1\right)\right| \\ &\quad + E\left|(\tilde{\beta}_{lp} - \hat{\beta}_{lp})\sum_{i=1}^{N} \frac{\tilde{m}_i}{N}\left(\frac{\delta_i}{\pi_i} - 1\right)\right| \\ &\quad + \left\{E\left(\sum_{i=1}^{N} \frac{(\hat{m}_i - \tilde{m}_i)^2}{N}\right)E\left(\hat{\beta}_{lp}^2 \sum_{i=1}^{N} \frac{(1 - \pi_i^{-1}\delta_i)^2}{N}\right)\right\}^{1/2}. \end{aligned} \quad \text{(A.9)}$$

First note that under assumptions (A1)–(A6), $\tilde{\beta}_{lp}$ is uniformly bounded and $\hat{\beta}_{lp}$ is uniformly bounded in s. Then, under (A1)–(A6) and using the fact that $\limsup_{\nu\to\infty} N^{-1}\sum_{i=1}^{N}(y_i - \tilde{m}_i\tilde{\beta}_{lp})^2 < \infty$ by lemma 2(iv) of Breidt and Opsomer (2000), the first term on the right side of (A.9) converges to 0 as $\nu \to \infty$, following the argument of theorem 1 of Robinson and Särndal (1983). By the Cauchy–Shwartz inequality, the second term on the right side of (A.9) is dominated by

$$\left\{E(\hat{\beta}_{lp} - \tilde{\beta}_{lp})^2 E\left[\sum_{i=1}^{N} \frac{\tilde{m}_i}{N}\left(\frac{\delta_i}{\pi_i} - 1\right)\right]^2\right\}^{1/2};$$

this converges to 0 because $\limsup_{\nu\to\infty} E(\hat{\beta}_{lp} - \tilde{\beta}_{lp})^2 < \infty$ for bounding arguments and, using the fact that $\limsup_{\nu\to\infty} \tilde{m}_i^2 < \infty$, the second factor converges to 0, following the argument of theorem 1 of Robinson and Särndal (1983). Let us consider the third term on the right side of equation (A.9),

$$E\left[\hat{\beta}_{lp}^2 \sum_{i=1}^{N} \frac{(1 - \pi_i^{-1}\delta_i)^2}{N}\right] \leq \left\{E(\hat{\beta}_{lp}^4)E\sum_{i=1}^{N} \frac{(1 - \pi_i^{-1}\delta_i)^4}{N^2}\right\}^{1/2}.$$

Because $\hat{\beta}_{nn}$ is uniformly bounded in s, $\limsup_{\nu\to\infty} E(\hat{\beta}_{lp}^4) < \infty$. Moreover,

$$\lim_{\nu\to\infty} E\frac{\sum_{i=1}^{N}(1 - \pi_i^{-1}\delta_i)^4}{N^2} = 0$$

for bounding arguments on π_i. Combining this with the fact that $\lim_{\nu\to\infty} N^{-1}\sum_{i=1}^{N}(\hat{m}_i - \tilde{m}_i)^2 = 0$ by lemma 4 of Breidt and Opsomer (2000), the third term in (A.9) converges to 0, and the theorem follows.

2. Asymptotic normality. From (A.8), it is clear that

$$\hat{\bar{Y}}_{lp}^{\mathrm{mc}} - \tilde{\bar{Y}}_{lp}^{\mathrm{mc}} = \sum_{i=1}^{N} \frac{\tilde{m}_i\tilde{\beta}_{lp} - \hat{m}_i\hat{\beta}_{lp}}{N}\left(\frac{\delta_i}{\pi_i} - 1\right).$$

The right side of this equation can be written as

$$(\tilde{\beta}_{lp} - \hat{\beta}_{lp})\sum_{i=1}^{N} \frac{\tilde{m}_i}{N}\left(\frac{\delta_i}{\pi_i} - 1\right) + \hat{\beta}_{lp}\sum_{i=1}^{N} \frac{\tilde{m}_i - \hat{m}_i}{N}\left(\frac{\delta_i}{\pi_i} - 1\right). \quad \text{(A.10)}$$

Now, because $\hat{m}_i - \tilde{m}_i = o_p(1)$ from lemma 4 of Breidt and Opsomer (2000), $\tilde{\beta}_{lp} - \hat{\beta}_{lp} = o_p(1)$ for an argument similar to that of Lemma A.4 here. Moreover, $N^{-1}\sum_{i=1}^{N} \tilde{m}_i(\delta_i/\pi_i - 1) = O_p(n^{-1/2})$ and $N^{-1} \times \sum_{i=1}^{N}(\tilde{m}_i - \hat{m}_i)(\delta_i/\pi_i - 1) = o_p(n^{-1/2})$ from the proof of theorem 2 of Breidt and Opsomer (2000). Therefore, the term in (A.10) is of order $o_p(n^{-1/2})$ and the argument follows.

Proof of Theorem 4

The result follows from arguments similar to those provided to prove Theorem 2 by noting that $r_i r_j = R_i R_j + o_p(1)$, because $\hat{m}_i - \tilde{m}_i = o_p(1)$ and $\hat{\beta}_{lp} - \tilde{\beta}_{lp} = o_p(1)$ from the proof of Theorem 3. Moreover, consistency of the Horvitz–Thompson variance estimator is guaranteed by assumptions (A6) and (A7); see, for example, the proof of theorem 3 of Breidt and Opsomer (2000).

[Received February 2003. Revised December 2004.]

REFERENCES

Barron, A. R. (1993), "Universal Approximation Bounds for Superpositions of a Sigmoidal Function," *IEEE Transactions on Information Theory*, 39, 930–945.

Breidt, F. J., and Opsomer, J. D. (2000), "Local Polynomial Regression Estimators in Survey Sampling," *The Annals of Statistics*, 28, 1026–1053.

Chambers, R. L. (1996), "Robust Case-Weighting for Multipurpose Establishment Surveys," *Journal of Official Statistics*, 12, 3–32.

——— (1998), "Weighting and Calibration in Sample Survey Estimation," in *Proceedings of a Conference on Statistical Science Honouring the Bicentennal of Stefano Franscini's Birth*, eds. C. Malaguerra, S. Morgenthaler, and E. Ronchetti, Basel: Birhauser-Verlag.

Chambers, R. L., Dorfman, A. H., and Wehrly, T. E. (1993), "Bias Robust Estimation in Finite Populations Using Nonparametric Calibration," *Journal of the American Statistical Association*, 88, 268–277.

Cybenko, G. (1989), "Approximation by Superpositions of a Sigmoidal Function," *Mathematics of Control Signals, and Systems*, 2, 303–314.
Deville, J. C., and Särndal, C. E. (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87, 376–382.
Di Ciaccio, A., and Montanari, G. E. (2001), "A Nonparametric Regression Estimator of a Finite Population Mean," in *Book of Short Papers*, *CLADAG 2001*, eds. Instituto di Statistica, Facoltà di Economia, Universitá degli Studi, Palermo, pp. 173–176.
Dorfman, A. H. (1992), "Nonparametric Regression for Estimating Totals in Finite Population," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 622–625.
Dorfman, A. H., and Hall, P. (1993), "Estimators of the Finite Population Distribution Function Using Nonparametric Regression," *The Annals of Statistics*, 21, 1452–1475.
EPA (2000), *Mid-Atlantic Highlands Streams Assessment*, EPA/903/R-00/015, Philadelphia: U.S. Environmental Protection Agency Region 3.
Friedman, J. H. (1994), "An Overview of Predictive Learning and Function Approximation," in *From Statistics to Neural Networks*, eds. V. Cherkassky, J. Friedman, and H. Wechsler, Berlin: Springer-Verlag.
Fuller, W. A. (1975), "Regression Analysis for Sample Survey," *Sankhyā*, Ser. C, 37, 117–132.
Fuller, W. A., and Isaki, C. T. (1981), "Survey Design Under Superpopulation Models," in *Current Topics in Survey Sampling*, eds. D. Krewski, J. N. K. Rao, and R. Platek, New York: Academic Press, pp. 199–226.
Funahashi, K. (1989), "On the Approximate Realization of Continuous Mappings by Neural Networks," *Neural Networks*, 2, 183–192.
Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2001), *The Elements of Statistical Learning—Data Mining, Inference, and Prediction*, New York: Springer-Verlag.
Hwang, J. T. G., and Ding, A. A. (1997), "Prediction Intervals for Artificial Neural Networks," *Journal of the American Statistical Association*, 92, 748–757.
Ingrassia, S., and Davino, C. (2002), *Reti Neuronali e Metodi Statistici*, Milano: Franco Angeli.
Isaki, C. T., and Fuller, W. A. (1982), "Survey Design Under the Superpopulation Models," *Journal of the American Statistical Association*, 77, 89–96.
Kott, P. S. (1990), "Estimating the Conditional Variance of a Design Consistent Regression Estimator," *Journal of Statistical Planning and Inference*, 24, 287–296.
Kuo, L. (1988), "Classic and Prediction Approaches to Estimating Distribution Functions From Survey Data," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 280–285.
Nordbotten, S. (1996), "Neural Network Imputation Applied to the Norwegian 1990 Population Census Data," *Journal of Official Statistics*, 12, 385–401.
Opsomer, J. D., Breidt, F. J., Moisen, G. G., and Kauermann, G. (2003), "Model-Assisted Estimation of Forest Resources With Generalized Additive Models," Preprint Series 03-05, Iowa State University, Dept. of Statistics.
Opsomer, J. D., and Miller, C. P. (2004), "Selecting the Amount of Smoothing in Nonparametric Regression Estimation for Complex Surveys," Preprint Series 04-17, Iowa State University, Dept. of Statistics.
Opsomer, J. D., Moisen, G. G., and Kim, J. Y. (2001), "Model-Assisted Estimation of Forest Resources With Generalized Additive Models," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.
Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge, U.K.: Cambridge University Press.
Robinson, P. M., and Särndal, C. E. (1983), "Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling," *Sankhyā*, Ser. B, 45, 240–248.
Särndal, C. E. (1980), "On π-Inverse Weighting versus Best Linear Unbiased Weighting in Probability Sampling," *Biometrika*, 67, 639–650.
——— (1996), "Efficient Estimators With Simple Variance in Unequal Probability Sampling," *Journal of the American Statistical Association*, 91, 1289–1300.
Särndal, C. E., Swensson, B., and Wretman, J. (1992), *Model-Assisted Survey Sampling*, New York: Springer-Verlag.
Thompson, M. E. (1997), *Theory of Sample Surveys*, London: Chapman & Hall.
Wu, C. F. J. (1981), "Asymptotic Theory of Nonlinear Least Squares Estimation," *The Annals of Statistics*, 9, 501–513.
Wu, C. (1999), "The Effective Use of Complete Auxiliary Information From Survey Data," unpublished doctoral dissertation, Simon Fraser University.
——— (2003), "Optimal Calibration Estimators in Survey Sampling," *Biometrika*, 90, 937–951.
Wu, C., and Sitter, R. R. (2001), "A Model-Calibration to Using Complete Auxiliary Information From Survey Data," *Journal of the American Statistical Association*, 96, 185–193.